WEEK 1: DATA CLEANING AND FEATURE ENGINEERING REPORT

AI GROUP-11
WEEK-1

Date:08.18.2025

Prepared By :

Vyshnavi Samudrala
Pranamya Praveen

1. Dataset Handling

Introduction

This report highlights the first steps taken to prepare Excelerate's learner dataset for analysis. In Week 1, the focus was on getting familiar with the data, cleaning it up, and creating a few new features. These steps are important because they make sure the dataset is accurate, consistent, and ready for deeper analysis in the coming weeks. By building this foundation early, we set ourselves up to uncover meaningful insights about learner engagement and eventually use the data for predictive modeling.

Data Description

The given dataset contains **non-identifying information about every learner who has ever created an account on the platform**. Each row represents a single user and includes details that give a well-rounded picture of the learner base.

Some of the key types of information include:

- **User ID** – a unique identifier for each learner.
- **Demographics** – details such as Date of Birth, Gender, City, State, and Country.
- **Education/Background** – academic major, level of study, and related fields.
- **Opportunity Information** – dates when a learner applied, as well as start and end dates for opportunities.
- **Engagement Details** – information such as time spent in an opportunity, completion status, and related metrics.

The dataset is broad and inclusive—it covers all learners, not just those who took part in specific opportunities. This makes it especially useful for identifying overall patterns and trends in engagement, while also allowing us to drill down into more detailed insights later on.

2. Data Cleaning Summary:

Dataset Size: 8,559 rows, 18 columns
Duplicates: Removed 1 duplicate row
Date Standardization: Converted all date fields (SignUp, Apply, Start, End, Entry Created, DOB) to YYYY-MM-DD format
Age: Recalculated from Date of Birth where valid; blanks kept if missing/unrealistic
Gender: Normalized to Male, Female, Other, Don't want to specify

Country: Standardized (e.g., US/USA → United States; UK → United Kingdom)
Institution Name: Trimmed and corrected spacing
Major: Removed junk entries, normalized common acronyms (e.g., CS → Computer Science), title-cased values
Opportunity Duration (Days): Recomputed when dates available; invalid values set to blank
Final Checks: Whitespace is trimmed, and consistent formatting applied across text fields
Remaining Missing Values:

- ➔ Opportunity Start Date: 4,436 (expected, many users not started)
- ➔ Apply Date: 201
- ➔ Opportunity Duration (Days): 527
- ➔ Current/Intended Major: 15
- ➔ Most other fields: 0–1 missing

Outcome: The dataset is cleaned, standardized, and ready for feature engineering and validation.

Issues Encountered and How I Resolved Them

- **Mixed Date Formats:** Some of the date fields were in different formats (for example, one record showed 06/14/2023 12:30:35 while another used 2023-01-05 05:29:16). This would have caused problems in analysis. I standardized all date fields into a consistent YYYY-MM-DD format to make them uniform.
- **Duplicate Records:** I noticed that the dataset had one duplicate row. Since duplicates can distort results, I removed it so that each record represented a unique learner.
- **Missing Values:** Several columns, like Opportunity Start Date and Institution Name, had missing values. For Opportunity Start Date, I realized many users had not yet started their opportunities, so I left those as blanks rather than filling them. For less critical fields, I retained blanks instead of dropping rows to avoid losing valuable data.
- **Inconsistent Text Entries:** Some fields had inconsistent naming, such as st. louis vs. Saint Louis or abbreviations like CS instead of Computer Science. I cleaned these up by standardizing cities and expanding majors into full names so the data was consistent across records.
- **Erroneous Inputs:** A few learners entered random or nonsensical text as their major (for example, strings of letters). I removed those invalid entries and kept only meaningful academic majors.

3. Feature Engineering

Feature engineering focused on transforming raw data into meaningful features to capture and analyze user engagement.

New Features Created

Age of Learner: Derived from Date of Birth to understand demographic engagement patterns.
Opportunity Duration: Computed as the difference between Opportunity Start Date and End Date to capture length of involvement.

Transformations Applied

- ➔ **Normalization:** Scaled numerical features like Age and Time in Opportunity for consistent analysis.
- ➔ **Encoding:** Applied one-hot encoding for categorical data such as Gender, State, and Country to make them usable in analysis.

Feature Examples

Example 1: Age of Users

- **Technique Used:** Date-based calculation.
- **Explanation:** Age was derived from the learner's Date of Birth column using the difference between today's date and the DOB.
- **Why it's useful:** This helps identify patterns in engagement across different age groups (e.g., whether younger learners are more engaged compared to older ones).
- **Illustration:** A learner born on *1998-06-15* was calculated as **26 years old**.

Example 2: Opportunity Duration

- ➔ **Technique Used:** Time difference calculation.
- ➔ **Explanation:** Duration was calculated by subtracting the Opportunity Start Date from the Opportunity End Date.
- ➔ **Why it's useful:** It provides insight into whether longer opportunities lead to more or less engagement.
- ➔ **Illustration:** An opportunity from *2023-01-01* to *2023-01-15* resulted in a **14-day duration**.

4. Data Validation

Once the cleaning and feature engineering steps were completed, the dataset was validated to make sure everything was accurate, consistent, and ready for analysis. This step was about double-checking that the data made sense, followed the right formats, and matched the expectations we set.

Validation Checks Performed

- **Duplicate Check**
  I confirmed that duplicate rows were removed and that every User ID is unique.
  ✅ *Result: The final dataset has no duplicates.*
- **Missing Values Review**
  I checked for missing values, especially in critical columns. Optional fields (like some demographic details) were left empty if learners hadn't provided them, while key columns were filled or verified.
  ✅ *Result: Columns like User ID, Date of Birth, and Opportunity Dates are complete.*
- **Format Consistency**
  I standardized date formats to `YYYY-MM-DD` and corrected inconsistencies in categorical data (for example, "St. Louis" vs. "Saint Louis").
  ✅ *Result: All dates now follow the same format and categories are consistent.*
- **Logical Validation**
  I verified that the new features made sense—for example, ages were positive and opportunity end dates came after start dates.
  ✅ *Result: All derived values are logical and fall within expected ranges.*
- **Outlier Review**
  I looked at extreme values (like unusually high ages or very long opportunity durations) and corrected or removed anything unrealistic.
  ✅ *Result: No unreasonable outliers remain in the dataset.*

Conclusion

Summary

In Week 1, our team focused on preparing the Excelerate dataset so it's ready for meaningful analysis. We started by getting familiar with the structure of the data and its key columns. From there, we cleaned the dataset by fixing missing values, removing duplicates, and standardizing entries that were inconsistent (like city names and random text in the major field).

We also added new features to give the dataset more depth, including **Age of Users** (calculated from their date of birth) and **Opportunity Duration** (the length of time between start and end dates). These additions make the dataset richer and more useful for understanding engagement. After validating everything for accuracy and consistency, we now have a dataset that's both clean and reliable.

Next Steps

Moving into Week 2, the focus will be on **Exploratory Data Analysis (EDA)**. Using the cleaned dataset, we'll look at patterns in learner engagement, check for seasonal or demographic trends, and highlight any unusual behaviors worth investigating. The goal is to turn the cleaned and structured data from Week 1 into insights that can guide predictive modeling and improve how Excelerate understands and supports its learners.